Identifying Language from Raw Speech --- An Application of Recurrent Neural Networks

Article · November 2001

Source: CiteSeer

CITATIONS
4

READS
76

4 authors, including:

Stan C. Kwasny
Washington University in St. Louis
47 PUBLICATIONS **476** CITATIONS

SEE PROFILE

Identifying Language from Raw Speech — An Application of Recurrent Neural Networks

Weilan Wu, Stan C. Kwasny, Barry L. Kalman, E. Maynard Engebretson
Department of Computer Science
Washington Unversity
St. Louis, MO 63130

Abstract

People can differentiate spoken languages without understanding them, and, in some sense, this differentiation can only be done without understanding the language. When we consider a multi-lingual person trying to understand an utterance spoken in one of the languages with which they are familiar, they will first decide which language this utterance belongs to, before trying to interpret it.

The language identification task is one example of high-level feature abstraction from raw speech. Speech samples are classified into categories according to what language was spoken. We conjecture that this classification can be performed reliably and in real time. To be successful, such a system should be speaker-independent as well as context-independent. A large amount of training is required to achieve a satisfactory level of performance.

We present a continuation of previous work (Kwasny et al., 1992) which introduces two important improvements to the system: (1) replacement of the non-recurrent, feed-forward network with a recurrent one, which is smaller, but still classifies correctly; (2) development of a frontend processor on a Next® workstation to facilitate sample recording and data acquisition. This is important for large-scale data acquisition, training, and testing of the network.

Introduction

Can languages be identified automatically from speech? People certainly can do this well. In an international setting, one might overhear parts of conversations in a variety of languages. Given the proper experience, identifying familiar languages can be done easily and fairly accurately. However, if speech samples are examined by the computer, can a system be built to reliably identify the language being spoken in real time? Certainly, acoustic (Hanley et al., 1966) and phonetic (Denes, 1963) differences exist among languages. Can these differences be leveraged in a system that performs this form of recognition?

How do we know, for example, when the same speaker speaks English or French? Under the right circumstances, people seem to be able to tell immediately, often not from exactly what is being said, but from broad characteristics of the speech. Spoken language is perceived on many levels. Listeners unconsciously notice many things about the speaker, such as accent, degree of excitement, the identity of the speaker, etc. These features can be very high level although often not consciously contemplated under ordinary circumstances by the listener. We further observe that humans communicate in speech through frequency, pitch, sequencing, and other prosodic and durational properties measurable in the signal. The complexity of interactions among these in the speech signal are impossible to capture in any simplistic model requiring the use of models such as Hidden Markov Models (HMM) and neural networks.

In preliminary work (Kwasny et al., 1992), we built a feed-forward network trained to distinguish between two languages: English and French. Only data from two speakers were used in training. The trained network showed the ability to differentiate between those two languages and some generalization capacity was observed. Data collected from a third speaker speaking English was used for testing, and the system did reliably identified the speech as English.

In order to scale up the system's performance and allow it to be trained on more speakers, we adopted a recurrent network

architecture, which is more suitable for processing temporal and sequential signals like speech. Our experiments substantiated some of the theory — the recurrent network was smaller, and its performance increased. We are in the process of developing a user interface on a Next® workstation to facilitate data collection and to verify the real-time claim.

This paper describes some initial evidence that, with the proper architecture and training, a neural network can achieve this capability. The applications of such a system range from long-distance telephone service, to a speech practice system for the deaf, to a frontend for language translation.

Background

There have been several studies that demonstrate the existence of statistically significant differences among spoken languages at the acoustic level (Hanley, et al., 1966; Atkinson, 1968) and also at the level of phonetic features (Denes 1963; Kucera & Monroe, 1968). Abe et al. (1990; 1991) have considered some of the differences in automatically converting a speaker's voice from one language into another. Since these differences are measurable at the low end of the speech chain, then it must be possible to exploit those differences to build a model that discriminates among languages as well as speakers.

House (1977) proposed a method of language identification which utilized a language structure component in conjunction with a statistical component. His approach was apparently hindered, at that time, by the lack of sufficient computing power to perform the necessary statistical procedures.

Similarly, we believe that statistical procedures are valuable in extracting certain high-level features. Being statistically based, the processing will naturally be resistant to noise and tolerant to some variation. We further assume that such processing can be demonstrated in real time. This assumption rules out the existence of a sophisticated language structure component and demands that intermediate levels of processing normally associated with speech understanding be finessed.

Recently, Muthusamy et al. (1990) has followed some of the suggestions made by House in examining this problem for four languages: American English, Japanese, Mandarin Chinese, and Tamil. They recorded six male and six female speakers each speaking 20 utterances in one of the languages. Four waveform and four spectral parameters were extracted and used by a neural network to segment and label the speech with one of 7 broad phonetic categories with 82.3% accuracy. The segmented speech was then used in a second network designed to classify by language. This proved to be 79.3% accurate in classifying the speech into one of four languages.

Our approach differs from theirs in several respects. We assume all processing can be conducted in real time and consequently cannot build sophisticated intermediate structures. Certainly some information is lost in mapping the waveform into discrete structures and this loss will have some effect on the success of the classification.

House proposed a statistical component, namely, an HMM, to capture the temporal (sequential) information within speech. Recurrent networks have many of the same features as HMMs. The recurrent net is a feedforward network with additional feedback (copy) links from hidden (and/or output) units to the input units. This kind of architecture is able to handle variable-length inputs and the sequential information among inputs. Since speech is a complex time-sequential process, the recurrent network has been viewed as the natural choice for modeling the temporal structure of speech signals. Bridle (1990) has shown some equivalence between the HMM and the recurrent neural network and a variety of hybrid HMM/network models have been proposed and evaluated (Austin et. al 1991).

Collection of Speech Samples

For our preliminary experiments, we collected speech samples from two bilingual speakers, one male and one female. British English was the native language of one speaker while the other speaker was native French. Both speakers were fluent in their non-native tongue. All recordings were made in an anechoic chamber resulting in 16-bit

<< Fig. 1 About Here >>

Figure 1: Speech Signal Preprocessing

samples at 24kHz. Five Band-Pass filters were used to separate the signal into bands which were low-pass filtered and decimated by a factor of 200. This process is illustrated in Figure 1.

Two samples of 12.5 second duration in each language, English and French, were collected for each speaker, for a total of 100 seconds of speech collected. Numerous overlapping samples of shorter duration were extracted from each collected sample by clipping segments of speech from the larger sample. Sequences of these shorter, overlapping segments were used to train and test the network.

Encouraged from our preliminary experiments, we are developing a user interface on a turbo Next® workstation, to provide on-line data acquisition and processing capabilities making it easier to collect speech samples and evaluate network training. As stated by Muthusamy et al (1990), ‘‘Despite several important applications of automatic language identification, this area has suffered from the absence of a standardized, public-domain database of languages.’’ We anticipate that scaling up to more languages and speakers will require large amounts of data.

In the user interface, speech samples can be recorded either in 8-bit format sampled at 8kHz through the built-in microphone or from an externally connected digital ear in 16-bit format sampled at 22-24kHz. Samples are transformed by the built-in CODEC (COder and DECoder) and fed to the DSP chip, which is programmed to perform the filtering and decimation algorithms shown in Figure 1. Data from the DSP can be divided into smaller samples and used as input to the neural network classifier (see Kwasny et al 1992).

Collecting samples is designed to be done in real-time, as compared with the time-consuming and unnatural method of using an anechoic chamber. Of course, recording in a computer room does introduce noise, but we expect training to overcome this problem.

Network Design

In our initial design, we succeeded in training a feed-forward network to correctly recognize bilingual speakers of English and French for virtually all samples. (Kwasny, et al., 1992). In the design discussed here, we use a simple recurrent network (Elman, 1990), in which the activation of hidden units are copied back as additional input unit activations, as shown in Figure 2.

Instead of using a 750 ms. duration, we use 400 ms. of overlapping speech segments. The shorter duration (and smaller network) is

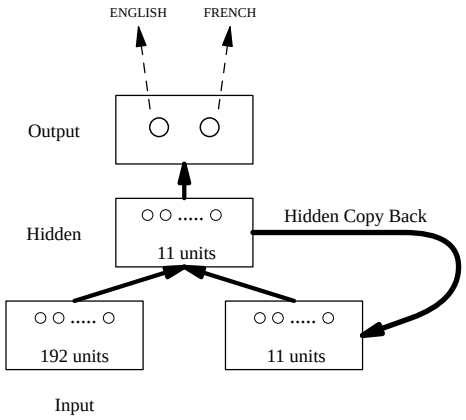

Figure 2: Recurrent Network Architecture

possible because the recurrent network is capable of encoding state. In the experiments, 400 ms. of speech results in 192 input units, compared with 360 units required for 750 ms. of speech.

The network has 11 hidden units (and therefore 11 copy-back units) and 2 output units, one for English and one for French. To permit even more samples, we divided the 12.5 second samples in two 6.25 second segments and moved the 400 ms. window along through each 6.25 second sample at intervals of 25ms. This resulted in sequences of overlapping speech segments which are used as training and testing sequences for the recurrent network.

In analyzing the data by bands, the middle or third band (8Hz − 16Hz) shares much with the adjacent bands. We decided to attempt to train the network from data further reduced by the elimination of band three. We successfully trained the network to approximately the same level without including band three. This training is faster since the network is smaller and there are fewer weights to adjust. This method of training was used for all the results reported here. But as more noisy recordings are made, we may need to include the third band, since introducing redundancy should improve reliability.

Results

Training did not succeed very well at first. Later, we applied singular value decomposition (svd) to the input patterns essentially reorienting the data to maximize orthogonality among the input unit activations (Kalman, et al., 1993). This technique permitted training to succeed. As each window is moved across the signal, the network decides if it contains an English or a French segment. Thus, the network is ‘‘voting’’ every 25 ms. and a simple majority vote determines the classification of the sample.

The speech samples of the two subjects were divided into training samples and testing samples. Each training sample was processed into 190 overlapping 400ms speech segments each of which produced 192 numeric values of frequency information across the four bands (48 samples of 4 bands). There are totally 235 samples in each 12.5 second sample, and the first 45 samples were used as a ‘‘warm-up’’ for the network to determine its context pattern. No votes are counted during this warm-up and no adjustments of the weights are performed during training. Training proceeded to settle at approximately 78.9 % correct on the test patterns.

If we consider the individual votes to be independent and apply the binomial theorem (see Kwasny et al., 1992), we can estimate the number of votes, or length of a speech sample, necessary to achieve a given level of identification accurary, based on the classification performance of the network. At a 78.9% correctness rate, less than 41 votes, or a speech sample of 1.75 second, will be sufficient to achieve a 99.5% accurary in language identification. Of course, this may be a slight underestimate, because of the independency assumption.

In a similar small experiment, we utilized the same data and trained the network to identify the gender of the two speakers to verify the model’s performance on other high-level feature extraction task. The results were even better than for language identifation, as expected. In this experiment, we used a 192+7-7-2 network, which achieved a 743/760 (97.8%) pattern classification correctness rate. Since this was so high, we did not test it with voting, which would undoubtedly pushed the performance toward perfection.

Discussion and Future Work

Recurrent networks have the potential to process time-ordered events, such as speech, but they are slow to train since representations in the copyback units must be literally invented since these are unsupervised units. With proper training, these representations emerge.

The improvement in accuracy shown by the recurrent network over the simple, feed-forward network should not be surprising. With the non-recurrent network, the limits of the window are also the limits of the network in its ability to corroborate surrounding information. A recurrent network might correlate

events disburse throughout the speech sample by encoding information as part of its distributed pattern.

While the work reported here is preliminary, the results are so encouraging that we are developing our user interface on the Next® further to test and evaluate these methods more thoroughly. Our goal is to complete a demonstration platform soon as proof that such processing can occur in real time and to enable several questions to be answered conveniently. Can the system be fooled by speech that sounds French, but is really nonsense? What happens if a bilingual speaker switches back and forth between French and English? How is performance affected if there is a great deal of background noise present?

To be useful, the system must be speaker-independent and context-independent. As more speakers and languages are attempted, more data will be necessary to train the system. We plan to continue to evaluate this architecture and apply it in extracting a variety of high-level features.

References

Abe, M., and Shikano, K. 1991. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *Journal of the Acoustic Society of America* 90: 76-82.

Abe, M.; Shikano, K.; and Kuwabara, H. 1990. Cross-language voice conversion. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 345-348.

Atkinson, K. 1968. Language identification from nonsegmental cues. *Journal of the Acoustic Society of America* 44, 378(A).

Austin, S.; Zavaliagkos G.; Makhoul J.; and Schwartz R., 1991. A Hybrid Continuous Speech Recognition System Using Segmental Neural Nets with Hidden Markov Models. Proceedings of the 1991 IEEE Workshop on Neural Networks for Signal Processing, IEEE Press, New York, NY, 1991, pp 347-356.

Bridle John S. 1990. Alpha-Nets: A Recurrent 'Neural' Network Architecture with a Hidden Markov Model Interpretation, *Speech Communication 9 (1990) 83-92* North-Holland.

Denes, P.B. 1963. On the statistics of spoken English. *Journal of the Acoustic Society of America* 35, 892-904.

Hanley, T.D., Snidecor, J.C., and Ringel, R.L. 1966. Some acoustic difference among languages. *Phonetica* 14, 97-107.

House, A.S., and Neuberg, E.P. 1977. Toward automatic identification of the language of an utterance. I. Preliminary methodological consideration. *Journal of the Acoustic Society of America* 62(3), 708-713.

Kwasny, S.C., Barry L. Kalman,Weilan Wu, and A. Maynard Engebretson, 1992. Identifying Language from Speech: An Example of High-Level, Statistically-Based Feature Extraction. In Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society, 909-913.

Kucera, H., and Monroe, G.K. 1968. A comparative quantitative phonology of Russian, Czech, and German. New York: American Elsevier.

Muthusamy, Y.K.; Cole, R.A.; and Gopalakrishnan, M. 1991. A segment-based approach to automatic language identification In Proceedings of the 1991 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada.